Task Similarity Aware Meta Learning: Theory-inspired Improvement on MAML

Pan Zhou*, **Yingtian Zou**†, **Xiaotong Yuan**‡, **Jiashi Feng**†, **Caiming Xiong***, **Steven C.H. Hoi***
*Salesforce Research, † National University of Singapore
† Nanjing University of Information Science & Technology
{pzhou, cxiong, shoi}@salesforce.com, zouyingt@comp.nus.edu.sg
xtyuan@nuist.edu.cn, elefjia@nus.edu.sg

Abstract

Few-shot learning ability is heavily desired for machine intelligence. By meta-learning a model initialization from training tasks with fast adaptation ability to new tasks, model-agnostic meta-learning (MAML) has achieved remarkable success in a number of few-shot learning applications. However, theoretical understandings on the learning ability of MAML remain absent yet, hindering developing new and more advanced meta learning methods in a principle way. In this work, we solve this problem by theoretically justifying the fast adaptation capability of MAML when applied to new tasks. Specifically, we prove that the learnt meta-initialization can quickly adapt to new tasks with only a few steps of gradient descent. This result, for the first time, explicitly reveals the benefits of the unique designs in MAML. Then we propose a theory-inspired task similarity aware MAML which clusters tasks into multiple groups according to the estimated optimal model parameters and learns group-specific initializations. The proposed method improves upon MAML by speeding up the adaptation and giving stronger few-shot learning ability. Experimental results on the few-shot classification tasks testify its advantages.

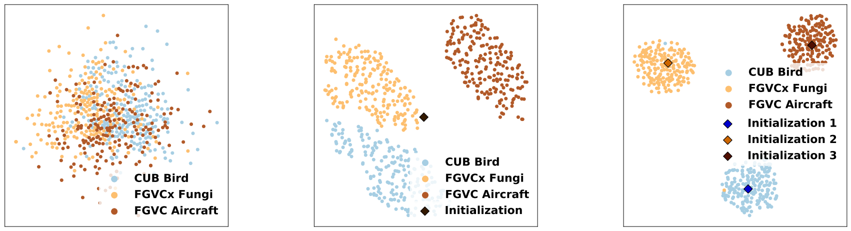
1 Introduction

Meta learning [1, 2, 3, 4], a.k.a. learning-to-learn [5], offers a new way to solve few-shot learning tasks via learning task-level knowledge. Specifically, at task level it trains a meta learner to extract task-shared knowledge from all the training tasks; then the meta learner is used to facilitate a task-specific model to learn a new task with only a small amount of data [6, 7, 8, 9, 10]. Among existing meta learning methods, model-agnostic meta-learning (MAML) [7] is a representative one because of its simplicity, generality and state-of-the-art performance [10, 11, 12]. It aims to learn a meta model from the observed tasks that could serve as a good initialization for task-specific models. Then given a test task, it only applies a few gradient descent steps on a few training samples for adapting the meta model to the test task, since the learnt initial model is desired to be close to the optimal models of the observed tasks and thus can be quickly adapted to new similar tasks.

Despite its remarkable success in practice [7, 13, 12], the theoretical understanding of MAML is still largely absent. Specifically, it is not clear *why MAML is able to generalize well in new tasks via merely taking a few steps of gradient descent on a small amount of data.* The answer to this question is important not only for justifying the fast adaptation capability of MAML, but also for inspiring new insights for algorithm improvement.

Contributions. In this work, we address the above fundamental question and contribute to derive some new results, insights and alternatives for MAML. Particularly, we provide rigorous theoretical analysis for its generalization behaviors. Inspired by our theory, we then propose a new alternative of MAML which is more effective for few-shot learning. Our main contributions are highlighted below.

4th Workshop on Meta-Learning at NeurIPS 2020, Vancouver, Canada.



(a) (b) (c)

Figure 1: Illustration of the learnt group structures by MAML and TSA-MAML on 5-shot 5-way learning task of a group-structured dataset with three sub-datasets, i.e. Aircraft [16], CUB Birds [17] and FGVCx Fungi [18]. One can observe indistinguishable sample features of tasks in (a) but well group-structured optimal model parameters of tasks learnt by MAML and TSA-MAML via 10 gradient descent steps from learnt initializations in (b) and (c) respectively. See details in Sec. 5.1.

Our first contribution is proving that in MAML, applying a few gradient descent steps on a small training dataset of a new task can achieve satisfactory generalization performance [14, 15] on its test data. Specifically, let $\boldsymbol{\theta}^*$ be the initialization learnt by MAML with meta model $f(\boldsymbol{\theta}, \boldsymbol{x})$ on the training tasks which are drawn from a task distribution $\mathcal{T}$. For a task T, let $\mathcal{L}_{D_T}(\boldsymbol{\theta}) = \frac{1}{K}\sum_{(\boldsymbol{x},\boldsymbol{y})\in D_T} \ell(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})$ denote its empirical risk on its training dataset D_T of size K. Then for any test task $T \sim \mathcal{T}$, we prove that its task-specific adapted parameter $\boldsymbol{\theta}_T^q = \boldsymbol{\theta}^* - \alpha\big[\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)\big]$ obtained by taking q gradient descent steps on its training data D_T has good performance on its test data $(\boldsymbol{x},\boldsymbol{y}) \sim T$, where $\boldsymbol{\theta}_T^1 = \boldsymbol{\theta}^* - \alpha\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*)$. Specifically, by defining population risk $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim T}\ell(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})$ on task T, we show the excess risk $\mathbb{E}_{T\sim\mathcal{T}}\mathbb{E}_{D_T}[\mathcal{L}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)]$ of $\boldsymbol{\theta}_T^q$, well measuring the testing performance, is upper bounded by $\mathcal{O}\big(\frac{\rho^q}{K} + \mathbb{E}_{T\sim\mathcal{T}}\mathbb{E}_{D_T}\big[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\big]\big)$, where the constant ρ is slightly larger than one, and $\boldsymbol{\theta}_T^*$ is the optimum of population risk $\mathcal{L}(\boldsymbol{\theta})$ on T. This result explicitly reveals the importance of the gradient step number q in MAML. Indeed, it suggests us to adapt the learnt initialization $\boldsymbol{\theta}^*$ to new task via a few gradient descent steps. See details in Sec. 3.2. Besides, we further upper bound $\mathbb{E}_{T\sim\mathcal{T}}\mathbb{E}_{D_T}\big[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\big]$ by $\frac{1}{2\alpha}\mathbb{E}_{T\sim\mathcal{T}}\big[\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_T^*\|_2^2\big]$, showing the smaller distance between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_T^*$ the smaller the excess risk. Meanwhile, as the learnt initialization $\boldsymbol{\theta}^*$ by MAML is often close to $\boldsymbol{\theta}_T^*$, our results can explain why MAML generalizes well to new tasks.

Inspired by our theory, we further develop the task similarity aware MAML (TSA-MAML) as a novel alternative to achieve faster adaptation to new tasks. As shown in Fig. 1 (a) and (b), though the samples in tasks are undistinguishable, the optimal model parameters estimated by MAML have remarkable group structures. So instead of learning one initialization for all tasks, TSA-MAML leverages task similarity to discover the group structures in the tasks by using a learner $\boldsymbol{\mathcal{A}}$ to measure task similarity in terms of the estimated task-specific model parameters. Then to facilitate the learning of new tasks, it learns multiple model initializations each of which corresponds to a group of similar tasks. Specifically, given a training task, TSA-MAML first uses the learner $\boldsymbol{\mathcal{A}}$ to predict its group membership and assign a group-specific initialization to it for few-shot training. Next, the initializations are in turn improved and become more group-specific. Consequently, as shown in Fig. 1 (c), the optimal model parameters of tasks in the same group are much closer to the group-specific initialization learnt by TSA-MAML than one common initialization learnt for all tasks by MAML. So TSA-MAML can adapt to new tasks more quickly and better under the few-shot learning setting. In this work, we implement the learner $\boldsymbol{\mathcal{A}}$ as the vanilla MAML and measure the task similarity according to the Euclidean distance between task-specific model parameters. We also theoretically show the superiority of TSA-MAML over MAML on learning new tasks. Extensive experimental results also well demonstrate the advantages of our approach on the few-shot learning problems.

2 Related Work

Meta learning has gained much attention recently because of its success in many applications [4, 7, 13, 19, 20, 21, 22]. The current methods can be divided as metric-based family [23, 9, 24, 25] that learns sample similarity metrics, memory-based family [26, 8, 27] that learns a fast adaptation algorithm via memory models [28], and optimization-based family [7, 6, 12, 10] that learns a model initialization for fast adaptation. Among them, optimization based methods are more preferable, thanks to its simplicity and effectiveness [7, 11, 29]. One representative method in this line is MAML [7] that learns a network initialization such that the network can adapt to a new task via a few gradient descent steps. Later, various variants are proposed to improve MAML [30, 12, 31]. Among them, HSML [31] considers the hierarchical parameter structures in tasks by learning task embeddings to measure task similarity. But it has two issues: (1) feature similarity cannot well reveal model parameter structures in tasks as shown in Fig. 1 and (2) learning similar embeddings for similar tasks is hard, as one cannot well align sample orders in tasks without global sample information (labels) and recurrent networks is sensitive to input orders. In contrast, we measure task similarity in the model parameter space and avoid the above issues. To handle multimodal task distribution, for a task T, MMAML [32] first learns its task embedding and then its task-specific parameter $\boldsymbol{\tau}$ which modulates meta-initialization $\boldsymbol{\theta}$ as inner-product initialization $\boldsymbol{\tau} \odot \boldsymbol{\theta}$ for T. It does not explicitly utilize task similarity as it still learns task-specific initialization. In contrast, we explore task structure by clustering similar tasks and learn group-specific initialization. Moreover, like [31], learning similar embeddings for similar tasks is hard. Besides, MMAML needs accurate task-specific parameter $\boldsymbol{\tau}$ to align with high-dimensional $\boldsymbol{\theta}$ to obtain accurate task initialization, increasing learning difficulty. TSA-MAML also differs from multi-task learning, *e.g.* [33, 34, 35], as TSA-MAML learns group-specific initialization with fast adaptation ability to new tasks, while the later directly learns task-specific optimal model.

The theoretical analysis of MAML is rarely investigated though heavily desired. Golmant [36] and Finn *et al.* [30] showed the convergence of MAML under strongly convex setting. In [37, 38], the convergence behavior of MAML on non-convex problems were studied. Saunshi *et al.* [39] analyzed the sample complexity for Reptile-alike algorithm [10] instead of MAML. The works [40, 41, 42, 43, 44] study the generalization performance of meta learning. But they focus on general meta learning methods and their results do not well reveal any unique property of MAML. For instance, they cannot explain why a few gradient descent steps on a few data in MAML is sufficient to obtain good testing performance. In contrast, by focusing on MAML itself, our theory well justifies this essential design in MAML. Besides, our results are more heuristic and directly derive a new MAML variant which leverages task similarity to facilitate new task learning and is well testified by experimental results.

3 Theoretical Analysis of MAML

Here we first briefly recall the formulation of MAML and then analyze the testing performance of the adapted task-specific model via a few gradient descent steps in MAML.

3.1 Formulation of MAML

MAML [7] is to learn a good initialization parameter $\boldsymbol{\theta}$ for a class of parameterized learner $f : \mathcal{X} \mapsto \mathcal{Y}$ (*e.g.* a classifier) such that for any task T drawn from a task distribution $\mathcal{T}$, its task-specific adapted parameter $\boldsymbol{\theta}_T$ via one gradient descent step from $\boldsymbol{\theta}$ on a small training dataset $D_T^{tr} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^K$ can perform well on its test dataset $D_T^{ts} = \{(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{y}}_i)\}_{i=1}^K$. Towards this goal, for each task $T \sim \mathcal{T}$, MAML optimizes the test loss of its adapted parameter $\boldsymbol{\theta}_T$ as follows

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{T\sim\mathcal{T}}\, \mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta} - \alpha \nabla \mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta})),$$

where $\mathcal{L}_{D_T}(\boldsymbol{\theta}_T) = \frac{1}{K}\sum_{(\boldsymbol{x},\boldsymbol{y})\in D_T} \ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})$ with $D_T = D_T^{tr}$ or D_T^{ts} is the empirical risk on the dataset D_T, and α is a learning rate. Here the function $\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})$ measures the discrepancy between the prediction $f(\boldsymbol{\theta}_T, \boldsymbol{x})$ and the ground truth $\boldsymbol{y}$, *e.g.* the cross-entropy loss in classification.

After learning the initialization $\boldsymbol{\theta}^*$, given a test task $T \sim \mathcal{T}$ with small training and test datasets D_T^{tr} and D_T^{ts} respectively, MAML adapts $\boldsymbol{\theta}^*$ to task T via a few gradient descent steps on D_T^{tr} and then tests the adapted parameter on D_T^{ts}. In spite of its impressive performance, there is no rigorously theoretical analysis of MAML that explicitly justifies effectiveness of a few gradient based adaptation. The following sections attempt to solve this issue by developing testing performance guarantees.

3.2 Testing Performance Analysis

Here we answer two questions: (1) why taking a few gradient descent steps on a few training data, MAML can achieve good performance on the test data; (2) how the learnt initialization benefits the learning of future tasks. Let $T \sim \mathcal{T}$ be any future task with K training samples $D_T = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^K$. Assume we run q gradient descent steps on the data D_T to obtain the adapted model $\boldsymbol{\theta}_T^q = \boldsymbol{\theta}^* - \alpha[\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1} \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)]$ for task T with learnt initialization $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_T^1 = \boldsymbol{\theta}^* - \alpha \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*)$. Let $\boldsymbol{\theta}_T^* \in \operatorname{argmin}_{\boldsymbol{\theta}_T} \{\mathcal{L}(\boldsymbol{\theta}_T) := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}) \sim T} [\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})]\}$ trained on all samples $(\boldsymbol{x}, \boldsymbol{y}) \sim T$ denote the optimal model parameter of the task $T \sim \mathcal{T}$. Before analysis, we first give necessary definitions which are fairly standard in the stochastic optimization [45, 46, 47, 48, 49] and the analysis of MAML [30, 36, 37, 38].

Definition 1 (Lipschitz continuity and smoothness)**.** *We say a function $g(\boldsymbol{\theta})$ is G-Lipschitz continuous if $\|g(\boldsymbol{\theta}_1) - g(\boldsymbol{\theta}_2)\|_2 \leq G\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ with a constant G. $g(\boldsymbol{\theta})$ is said to be L_s-smooth if $\|\nabla g(\boldsymbol{\theta}_1) - \nabla g(\boldsymbol{\theta}_2)\|_2 \leq L_s\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ with a constant L_s.*

Then we formally state our results in Theorem 1 which shows the role of q and the benefits of initialization $\boldsymbol{\theta}^*$ on reducing the excess risk $\mathsf{ER}(\boldsymbol{\theta}_T^q) = \mathbb{E}_{T\sim\mathcal{T}}\mathbb{E}_{D_T}\big[\mathcal{L}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\big]$. As $\mathsf{ER}(\boldsymbol{\theta}_T^q)$ evaluates the loss difference $[\ell(f(\boldsymbol{\theta}_T^q, \boldsymbol{x}), \boldsymbol{y}) - \ell(f(\boldsymbol{\theta}_T^*, \boldsymbol{x}), \boldsymbol{y})]$ on all samples $(\boldsymbol{x}, \boldsymbol{y}) \sim T$ and all tasks $T \sim \mathcal{T}$, it can well measure the testing performance of the adapted parameter $\boldsymbol{\theta}_T^q$ [14, 15].

Theorem 1. *(**Testing Performance Analysis**) Suppose $\ell(f(\boldsymbol{\theta}, \boldsymbol{x}), \boldsymbol{y})$ is G-Lipschitz continuous w.r.t. the parameter $\boldsymbol{\theta}$. We also assume $\ell(f(\boldsymbol{\theta}, \boldsymbol{x}), \boldsymbol{y})$ is L_s-smooth w.r.t. $\boldsymbol{\theta}$ and α obeys $\alpha \leq \frac{1}{L_s}$. By setting $\rho = 1 + 2\alpha L_s$, then for any $T \sim \mathcal{T}$ and $D_T = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^K \sim T$, we have*

$$\mathit{ER}(\boldsymbol{\theta}_T^q) \overset{(a)}{\leq} \frac{2G^2(\rho^q - 1)}{KL_s} + \mathbb{E}_{T\sim\mathcal{T}}\mathbb{E}_{D_T}\big[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\big] \overset{(b)}{\leq} \frac{2G^2(\rho^q - 1)}{KL_s} + \frac{1}{2\alpha}\mathbb{E}_{T\sim\mathcal{T}}\big[\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_T^*\|_2^2\big].$$

See its proof in Appendix B.2. From the first inequality (a) in Theorem 1, one can observe that the excess risk $\mathsf{ER}(\boldsymbol{\theta}_T^q)$ of the task-specific adapted model $\boldsymbol{\theta}_T^q$ for task T is determined by two factors, i.e., the training sample number K for each task and the expected loss distance $\mathbb{E}_{T\sim\mathcal{T}}\mathbb{E}_{D_T}\big[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\big]$ between the adapted parameter $\boldsymbol{\theta}_T^q$ provided by MAML and the optimal model $\boldsymbol{\theta}_T^*$ for task T. Obviously, the larger training sample number K is, the smaller the first term in the upper bound is. Besides, the closer $\boldsymbol{\theta}_T^q$ is to $\boldsymbol{\theta}_T^*$, the better task-specific parameter $\boldsymbol{\theta}_T^q$ with smaller excess risk.

From the results, one way to reduce the loss $\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q)$ is to increase the number q of gradient descent steps for adaptation which however also increases the first term in the upper bound, as ρ is often slightly larger than one since we often use a small learning rate α. To trade-off the first and second terms, q should not be large. This is because the second term $\mathbb{E}_{T\sim\mathcal{T}}\mathbb{E}_{D_T}\big[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\big]$ would decrease very fast at the first several iterations but reduce slowly along more optimization iterations, especially for small datasets (see the illustrations in Fig. 5 in Appendix A.3), while the first term always exponentially increases. This explains why MAML often adapts the learnt initialization $\boldsymbol{\theta}^*$ to new tasks via a few gradient descent steps, and also provides new insights to set step number q.

The second inequality (b) in Theorem 1 justifies the benefits of the learnt initialization $\boldsymbol{\theta}^*$ to the testing performance. Specifically, Theorem 1 shows the smaller distance between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_T^*$, the smaller excess risk. Intuitively, if $\boldsymbol{\theta}^*$ is close to $\boldsymbol{\theta}_T^*$, the task-specific adapted parameter $\boldsymbol{\theta}_T^q$ would be close to $\boldsymbol{\theta}_T^*$, guaranteeing good testing performance of $\boldsymbol{\theta}_T^q$ on its corresponding task $T \sim \mathcal{T}$. Fortunately, empirical results of MAML show that a few gradient steps from $\boldsymbol{\theta}^*$ can provide good performance for test task $T \sim \mathcal{T}$, indicating small distance $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_T^*\|_2^2$.

4 Task Similarity Aware MAML

Theorem 1 shows that if one hopes to achieve good testing performance, the learnt initialization $\boldsymbol{\theta}^*$ should be close to the optimal model parameter $\boldsymbol{\theta}_T^*$ of any task $T \sim \mathcal{T}$, i.e. small distance $\mathbb{E}_{T\sim\mathcal{T}}[\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_T^*\|_2^2]$. One natural way to further reduce this distance is to learn multiple initializations $\{\boldsymbol{\theta}_i^*\}_{i=1}^m$ and select a correct initialization $\boldsymbol{\theta}_{i_T}^* = \boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i^*\}_{i=1}^m, T)$ from $\{\boldsymbol{\theta}_i^*\}_{i=1}^m$ for a specific task T such that $\mathbb{E}_{T\sim\mathcal{T}}[\|\boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i^*\}_{i=1}^m, T) - \boldsymbol{\theta}_T^*\|_2^2]$ is small. Here given a task T, the learner $\boldsymbol{\mathcal{A}}$ assigns it into one of the m groups according to the similarity between T and the tasks in each group such that the optimal model parameter $\boldsymbol{\theta}_T^*$ of T is close to the initialization $\boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i^*\}_{i=1}^m, T)$ shared by the tasks in the same group. Here we focus on a general learner $\boldsymbol{\mathcal{A}}$ and provide one effective approach to implement it below. Towards this goal, we propose *task similarity aware MAML* (TSA-MAML):

$$\min_{\{\boldsymbol{\theta}_i\}_{i=1}^m, \boldsymbol{\mathcal{A}}} \ \mathbb{E}_{T\sim\mathcal{T}} \mathcal{L}_{D_T^{ts}}(\boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i\}_{i=1}^m, T) - \alpha \nabla \mathcal{L}_{D_T^{tr}}(\boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i\}_{i=1}^m, T))).$$

Intuitively, this model aims at using the learner $\boldsymbol{\mathcal{A}}$ to cluster tasks $T \sim \mathcal{T}$ into m groups according to their similarity in terms of their optimal model parameter estimation such that the tasks in each group are sufficiently close to a common initialization. Then based on Theorems 1, we derive the testing performance bound of TSA-MAML. Let $\{\boldsymbol{\theta}_i^*\}_{i=1}^m$ be the learnt multiple initializations, $\bar{\boldsymbol{\theta}}_T^* = \boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i^*\}_{i=1}^m, T)$ be the assigned initialization for task T, and $\boldsymbol{\theta}_T^q$ be the adapted parameter $\boldsymbol{\theta}_T^q = \bar{\boldsymbol{\theta}}_T^* - \alpha\big[\nabla\mathcal{L}_{D_T}(\bar{\boldsymbol{\theta}}_T^*) + \sum_{t=1}^{q-1}\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)\big]$ for task T with $\boldsymbol{\theta}_T^1 = \bar{\boldsymbol{\theta}}_T^* - \alpha\nabla\mathcal{L}_{D_T}(\bar{\boldsymbol{\theta}}_T^*)$. $\boldsymbol{\theta}_T^*$ is the optimal model parameter of the population risk $\mathcal{L}(\boldsymbol{\theta}_T) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim T}\left[\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})\right]$ on task T. Then we state our results in Corollary 1 with proof in Appendix C.1.

Corollary 1. *With the same assumptions in Theorem 1 and $\rho = 1 + 2\alpha L_s$, for any $T \sim \mathcal{T}$ and $D_T = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^K \sim T$, the expected excess risk $\mathsf{ER}(\boldsymbol{\theta}_T^q)$ and the population gradient $\mathsf{EPG}(\boldsymbol{\theta}_T^q)$ satisfies*

$$\mathsf{ER}(\boldsymbol{\theta}_T^q) \leq \frac{\tau_1}{KL_s} + \frac{1}{2\alpha}\mathbb{E}_{T\sim\mathcal{T}}\big[\|\boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i^*\}_{i=1}^m, T) - \boldsymbol{\theta}_T^*\|_2^2\big] \qquad \text{where } \tau_1 = 2G^2(\rho^q - 1).$$

Corollary 1 shows that if the learner $\boldsymbol{\mathcal{A}}$ can assign the task $T \sim \mathcal{T}$ into a correct group with a small distance $\mathbb{E}_{T\sim\mathcal{T}}[\|\boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i^*\}_{i=1}^m, T) - \boldsymbol{\theta}_T^*\|_2^2]$, TSA-MAML would be expected to have smaller expected excess risk $\mathsf{ER}(\boldsymbol{\theta}_T^q)$ and thus better testing performance than MAML. This can be intuitively understood: by grouping the tasks $T \sim \mathcal{T}$ into m clusters such that the tasks in the same group have similar optimal model parameters and by learning a group-specific shared initialization for each group, the optimal model parameters of tasks in a group will be much closer to the group-specific shared initialization learnt by TSA-MAML than a common initialization learnt for all tasks $T \sim \mathcal{T}$ in MAML. Accordingly, TSA-MAML requires less samples to adapt to new tasks and thus achieves better testing performance.

Algorithm 1 Meta Framework for TSA-MAML

Input: learning rates α and β, task distribution $\mathcal{T}$.
Initialization: initialize $\{\boldsymbol{\theta}_i^0\}_{i=1}^m$ via the vanilla MAML and k-means based approach.
for $t = 0, \cdots, S-1$ **do**
  sample a task mini-batch $\mathcal{S}^t = \{T_i\}_{i=1}^s$ as $T_i \sim \mathcal{T}$.
  for task T_i in $\mathcal{S}^t$ **do**
    set initialization $\boldsymbol{\theta}_{i_{T_i}} = \boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i^t\}_{i=1}^m, T_i)$ for T_i.
    compute gradient $\nabla\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}_{i_{T_i}})$.
    update task-specific parameter $\boldsymbol{\theta}_{T_i}$ as $\boldsymbol{\theta}_{T_i} = \boldsymbol{\theta}_{i_{T_i}} - \alpha\nabla\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}_{i_{T_i}})$ for task T_i.
  end for
  update $\{\boldsymbol{\theta}_i^{t+1}\}_{i=1}^m$ as follows:
  $\{\boldsymbol{\theta}_i^{t+1}\}_{i=1}^m = \{\boldsymbol{\theta}_i^t\}_{i=1}^m - \beta\sum_{T_i\sim\mathcal{T}}\nabla_{\{\boldsymbol{\theta}_i^t\}_{i=1}^m}\mathcal{L}_{D_{T_i}^{ts}}(\boldsymbol{\theta}_{T_i})$.
end for
Output: $\{\boldsymbol{\theta}_i^S\}_{i=1}^m$

Implementation. The key for implementing TSA-MAML is to design the learner $\boldsymbol{\mathcal{A}}$ which assigns a task T into a correct group such that its optimal model parameter is close to the initialization of the group. Here we implement $\boldsymbol{\mathcal{A}}$ as follows. Firstly, we train vanilla MAML and obtain the initialization $\boldsymbol{\theta}^*$ for all tasks $T \sim \mathcal{T}$. Then we use vanilla MAML with initialization $\boldsymbol{\theta}^*$ to compute the estimated optimal parameters $\{\bar{\boldsymbol{\theta}}_{T_i}\}_{i=1}^n$ of sufficient sampled tasks $\{T_i\}_{i=1}^n$ and perform k-means [50] on $\{\bar{\boldsymbol{\theta}}_{T_i}\}_{i=1}^n$ to cluster them into m groups $\{\boldsymbol{\mathcal{G}}_i\}_{i=1}^m$. See the experimental settings of n and m in Sec. 5. Next, we initialize each group-specific initialization $\boldsymbol{\theta}_i^0$ by averaging the model parameters $\{\bar{\boldsymbol{\theta}}_{T_i}\}_{i\in\boldsymbol{\mathcal{G}}_i}$. Finally, for training, given a task T, we also first use vanilla MAML with initialization $\boldsymbol{\theta}^*$ to compute its estimated optimum $\bar{\boldsymbol{\theta}}_T$, and then find a group $\boldsymbol{\mathcal{G}}_i$ such that the group-specific initialization $\boldsymbol{\theta}_i$ has a smallest Euclidean distance to $\bar{\boldsymbol{\theta}}_T$. In this way, we can use task T to update the initialization $\boldsymbol{\theta}_i$ for group $\boldsymbol{\mathcal{G}}_i$ like MAML. Note, we measure the task similarity in the model parameter space instead of the task feature space (sample feature) which measures the similarity more accurately, since task features cannot well reveal the group structures of the optimal models of tasks which is illustrated by Fig. 1 and will be discussed in Sec. 5.1 with more details. See detailed algorithm in Algorithm 1.

5 Experiments

5.1 Evaluation on Group-Structured Data

Datasets. We investigate whether TSA-MAML can leverage the task similarity to discover task-group structures and further learn group-specific initializations. We randomly sample each training/test task from one of the three datasets, i.e. Aircraft dataset [16], CUB Birds [17] and FGVCx-Fungi dataset [18]. As each dataset only contains one category, *e.g.* birds, the tasks drawn from each dataset should have similar optimal model parameters, indicating remarkable group structures in these optimal model parameters as illustrated by Fig. 1. Accordingly, discovering these group structures and learning group-specific initializations can benefit new task learning. Similarly, we construct the second group-structured dataset which contains Stanford Car [51], CUB Birds [17] and FGVCx-Fungi [18]. Like conventional setting, each sub-dataset, *e.g.* CUB Birds, contains meta-training, meta-validation and meta-test classes which is specified in [31].

Results. Table 1 shows that TSA-MAML achieves the best performance over other state-of-the-arts. Specifically, on the first group-structured dataset (Aircraft + Birds + Fungi), TSA-MAML respectively makes about 2.95%, 2.84% and 3.23% improvements on the three sub-dataset (from left to right). It also brings about 4.00% improvement for the overall accuracy. Similarly, for the second group-structured dataset (Car + CUB Birds + Fungi), TSA-MAML also outperforms others on all three sub-datasets and averagely improves by about 2.36%. Compared with the approaches learning one common initialization, *e.g.* MAML and Reptile, TSA-MAML leverages task similarity in the model parameter space to discover the group structures in the tasks and learns group-specific initializations to facilitate the learning of new tasks, boosting the performance.

Fig. 2 further reports the usage frequency of the multiple initializations learnt by TSA-MAML when testing new tasks. After learning three initializations, we sample 1,000 test tasks from each sub-dataset of the group-structured dataset, and then assign one initialization for each test task by first using MAML to find its approximate optimal model $\boldsymbol{\theta}_T$ and selecting a learnt initialization with smallest distance to $\boldsymbol{\theta}_T$. The values in the (i, j)-th grid in Fig. 2 denotes the frequency that TSA-MAML assigns the i-th learnt initialization to the tasks from the j-th sub-dataset. From these results in Fig. 2, one can observe that in most cases, TSA-MAML assigns the same learnt initialization for the tasks from the same sub-dataset. This well demonstrates that TSA-MAML has leveraged the task similarity and thus can well learn the group structures in the tasks, explaining the superiority over state-of-the-arts.

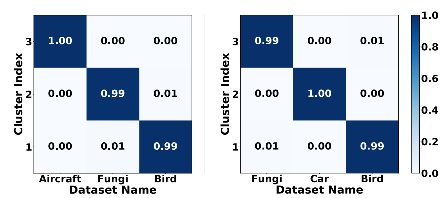


Figure 2: Usage frequency of multiple initializations in TSA-MAML on new tasks.

5.2 Evaluation on Real Data

We evaluate TSA-MAML on two benchmarks, tieredImageNet [52] and CIFARFS [53]. From Table 2 (tieredImageNet) and Table 3 (CIFARFS) in Appendix A, one can observe that TSA-MAML consistently outperforms the compared methods. Specifically, on tieredImageNet, it averagely improves by about 1.68% and 1.20% on the four test cases under non-transduction and transduction cases. Similarly, on CIFARFS, TSA-MAML respectively brings about 1.91% and 1.55%, 1.29% average improvements on the four test cases under non-transduction and transduction cases. These results demonstrate the advantages of TSA-MAML. Besides, compared with MAML, TSA-MAML

Table 1: Classification accuracy ($\%$) of the compared approaches on the 5-shot 5-way few-shot learning tasks in the two group-structured datasets (600 test episodes with 95% confidence intervals).

	Aircraft + CUB Bird + FGVCx Fungi				Stanford Car + CUB Bird + FGVCx Fungi			
	aircraft	bird	fungi	average	car	bird	fungi	average
Reptile [10]	60.46±0.68	71.96±0.79	51.71±0.84	61.38	43.64±0.64	69.63±0.78	52.06±0.85	55.11
HSML [31]	69.89±0.90	68.99±1.01	53.63±1.03	64.17	48.19±0.93	71.20±0.97	53.48±1.08	57.62
MMAML [32]	56.02±0.63	68.33±0.82	53.44±0.76	59.26	34.97±0.46	64.83±0.80	53.33±0.77	51.04
FOMAML [7]	49.60±0.98	69.53±0.95	47.56±0.83	55.56	34.20±0.72	68.50±0.78	46.66±0.89	49.79
MAML [7]	67.82±0.65	70.55±0.77	53.20±0.82	63.86	47.67±0.70	68.64±0.82	53.43±0.89	56.25
TSA-MAML	**72.84±0.63**	**74.80±0.76**	**56.86±0.87**	**68.17**	**50.01±0.65**	**73.92±0.80**	**56.03±0.87**	**59.98**

Table 2: Few-shot classification accuracy ($\%$) of the compared approaches on the tieredImageNet dataset. The reported accuracies are averaged over 600 test episodes with 95% confidence intervals.

method	1-shot 5-way	5-shot 5-way	1-shot 10-way	5-shot 10-way
Matching Net [9]	34.95 ± 0.89	43.95 ± 0.85	22.46 ± 0.34	31.19 ± 0.30
Meta-LSTM [6]	33.71 ± 0.76	46.56 ± 0.79	22.09 ± 0.43	35.65 ± 0.39
Reptile [10]	$\mathbf{49.12 \pm 0.43}$	65.99 ± 0.75	31.79 ± 0.28	47.82 ± 0.30
HSML [31]	47.36 ± 0.84	66.16 ± 0.78	33.39 ± 0.57	51.53 ± 0.55
MMAML [32]	44.82 ± 0.46	61.47 ± 0.49	30.42 ± 0.37	48.92 ± 0.29
FOMAML [7]	48.01 ± 1.74	64.07 ± 1.72	30.31 ± 1.12	46.54 ± 1.24
MAML [7]	48.50 ± 1.83	65.93 ± 1.78	32.41 ± 1.23	48.81 ± 1.32
TSA-MAML	48.82 ± 0.88	$\mathbf{67.82 \pm 0.72}$	$\mathbf{34.48 \pm 0.56}$	$\mathbf{52.26 \pm 0.55}$
Reptile + Transduction [10]	51.06 ± 0.45	66.30 ± 0.78	33.79 ± 0.29	51.27 ± 0.31
HSML + Transduction [31]	48.82 ± 0.86	66.74 ± 0.76	34.63 ± 0.55	51.47 ± 0.54
MMAML + Transduction [32]	48.52 ± 0.47	64.39 ± 0.47	33.69 ± 0.35	50.90 ± 0.29
FOMAML + Transduction [7]	50.12 ± 1.82	67.43 ± 1.80	31.53 ± 1.08	49.99 ± 1.36
MAML + Transduction [7]	50.48 ± 1.81	68.06 ± 1.75	34.25 ± 1.19	51.69 ± 1.33
TSA-MAML + Transduction	$\mathbf{52.03 \pm 0.86}$	$\mathbf{68.97 \pm 0.74}$	$\mathbf{35.78 \pm 0.58}$	$\mathbf{52.50 \pm 0.56}$

respectively makes about 1.44% and 1.73% average improvements on tieredImageNet and CIFARFS. These observations further confirm our theories in Sec. 3.2.

6 Conclusion

In this work, we theoretically justify the effectiveness of a few gradient based adaptation and the benefits of the learnt initialization for fast adaptation. Then we propose TSA-MAML as a new variant of MAML which leverages the task-similarity via learning shared initialization for similar tasks to facilitate learning new tasks. Experimental results demonstrate the superiority of TSA-MAML.

References

[1] J. Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta hook*. PhD thesis, Technische Universität München, 1987.

[2] D. Naik and R. Mammone. Meta-neural networks that learn by learning. In *Int'l Joint Conf. Neural Networks* , pages 437–442, 1992.

[3] Y. Bengio, S. Bengio, and J. Cloutier. Learning a synaptic learning rule. In *Int'l Joint Conf. Neural Networks* , 1990.

[4] P. Zhou, X. Yuan, H. Xu, S. Yan, and J. Feng. Efficient meta learning via minibatch proximal update. In *Proc. Conf. Neural Information Processing Systems*, 2019.

[5] S. Thrun and L. Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

[6] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *Int'l Conf. Learning Representations*, 2017.

[7] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. Int'l Conf. Machine Learning*, pages 1126–1135, 2017.

[8] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In *Proc. Int'l Conf. Machine Learning*, pages 1842–1850, 2016.

[9] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra. Matching networks for one shot learning. In *Proc. Conf. Neural Information Processing Systems*, pages 3630–3638, 2016.

[10] A. Nichol and J. Schulman. Reptile: a scalable meta-learning algorithm. *arXiv preprint arXiv:1803.02999*, 2, 2018.

[11] A. Antoniou, H. Edwards, and A. Storkey. How to train your MAML. In *Int'l Conf. Learning Representations*, 2019.

[12] Z. Li, F. Zhou, F. Chen, and H. Li. Meta-SGD: Learning to learn quickly for few-shot learning. In *Proc. Conf. Neural Information Processing Systems*, 2017.

[13] Y. Duan, J. Schulman, X. Chen, P. Bartlett, I. Sutskever, and P. Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

[14] P. Zhou and J. Feng. Understanding generalization and optimization performance of deep cnns. In *Proc. Int'l Conf. Machine Learning*, 2018.

[15] P. Zhou and J. Feng. Empirical risk landscape analysis for understanding deep neural networks. In *Int'l Conf. Learning Representations*, 2018.

[16] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[17] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[18] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. 2018 FGCVx fungi classification challenge. *fungi-challenge-fgvc-2018*, 2018.

[19] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. Meta-learning with temporal convolutions. *arXiv preprint arXiv:1707.03141*, 2(7), 2017.

[20] F. Sung, L. Zhang, T. Xiang, T. Hospedales, and Y. Yang. Learning to learn: Meta-critic networks for sample efficient learning. *arXiv preprint arXiv:1706.09529*, 2017.

[21] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, and P. Manzagol. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.

[22] H. Yao, X. Wu, Z. Tao, Y. Li, B. Ding, R. Li, and Z. Li. Automated relational meta-learning. *arXiv preprint arXiv:2001.00745*, 2020.

[23] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

[24] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. Torr, and T. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[25] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Proc. Conf. Neural Information Processing Systems*, pages 4077–4087, 2017.

[26] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[27] T. Munkhdalai and H. Yu. Meta networks. In *Proc. Int'l Conf. Machine Learning*, pages 2554–2563, 2017.

[28] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[29] M. Khodak, M. Balcan, and A. Talwalkar. Provable guarantees for gradient-based meta-learning. *arXiv preprint arXiv:1902.10644*, 2019.

[30] C. Finn, A. Rajeswaran, S. Kakade, and S. Levine. Online meta-learning. *arXiv preprint arXiv:1902.08438*, 2019.

[31] H. Yao, Y. Wei, J. Huang, and Z. Li. Hierarchically structured meta-learning. In *Proc. Int'l Conf. Machine Learning*, 2019.

[32] R. Vuorio, S. Sun, H. Hu, and J. Lim. Multimodal model-agnostic meta-learning via task-aware modulation. In *Proc. Conf. Neural Information Processing Systems*, pages 1–12, 2019.

[33] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *Proc. Int'l Conf. Machine Learning*, volume 2, page 4, 2011.

[34] A. Kumar and H. Daume. Learning task grouping and overlap in multi-task learning. In *Proc. Int'l Conf. Machine Learning*, 2013.

[35] A. Pentina, V. Sharmanska, and C. Lampert. Curriculum learning of multiple tasks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 5492–5500, 2015.

[36] N. Golmant. On the convergence of model-agnostic meta-learning. *http://noahgolmant.com/writings/maml.pdf*, 2019.

[37] A. Fallah, A. Mokhtari, and A. Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. *arXiv preprint arXiv:1908.10400*, 2019.

[38] K. Ji, J. Yang, and Y. Liang. Multi-step model-agnostic meta-learning: Convergence and improved algorithms. *arXiv preprint arXiv:2002.07836*, 2020.

[39] N. Saunshi, Y. Zhang, M. Khodak, and S. Arora. A sample complexity separation between non-convex and convex meta-learning. *arXiv preprint arXiv:2002.11172*, 2020.

[40] J. Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[41] A. Maurer. Algorithmic stability and meta-learning. *J. of Machine Learning Research*, 6(Jun):967–994, 2005.

[42] R. Amit and R. Meir. Meta-learning by adjusting priors based on extended PAC-bayes theory. In *Proc. Int'l Conf. Machine Learning*, 2018.

[43] K. Mikhail, B. Maria-Florina, and T. Ameet. Adaptive gradient-based meta-learning methods. In *Proc. Conf. Neural Information Processing Systems*, 2019.

[44] S. Du, W. Hu, S. Kakade, J. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[45] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[46] P. Zhou, X. Yuan, and J. Feng. Faster first-order methods for stochastic non-convex optimization on riemannian manifolds. 2019.

[47] P. Zhou, X. Yuan, and J. Feng. New insight into hybrid stochastic gradient descent: Beyond with-replacement sampling and convexity. In *Proc. Conf. Neural Information Processing Systems*, 2018.

[48] P. Zhou and X. Tong. Hybrid stochastic-deterministic minibatch proximal gradient: Less-than-single-pass optimization with nearly optimal generalization. In *Proc. Int'l Conf. Machine Learning*, 2020.

[49] P. Zhou, J. Feng, C. Ma, C. Xiong, S. Hoi, and W. E. Towards theoretically understanding why SGD generalizes better than ADAM in deep learning. In *Proc. Conf. Neural Information Processing Systems*, 2020.

[50] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Symposium on Math, Statistics, and Probability*, page 281–297, 1967.

[51] J. Krause, M. Stark, J. Deng, and F. Li. 3D object representations for fine-grained categorization. In *Int'l IEEE Workshop on 3D Representation and Recognition*, 2013.

[52] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. Tenenbaum, H. Larochelle, and R. Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[53] L. Bertinetto, J. Henriques, P. Torr, and A. Vedaldi. Meta-learning with differentiable closed-form solvers. In *Int'l Conf. Learning Representations*, 2019.